CAPACITY MANAGEMENT PROCEDURES

Subject: Scientific Data Storage and Access (SDSA) Area

Author: **Gene Oleynik**

Date: **10/19/2015**

The purpose of this document is to document the procedures used to manage service and component capacity.

In this document, "customer" refers to a Fermilab experiment or a supported scientific organization.

CAPACITY MANAGEMENT SUMMARY

The following table summarizes the capacity plan for the Backup and Restore service:

Capacity Metric	Capacity Requirement	Predicted Growth + Timescale	Capacity Threshold	Threshold Response Strategy/Tuning (Action to Be Take Upon Reaching Threshold(s), includes any tuning or demand management strategies)
Tape Slots	~3500/yr	3500 (2000 CMS + 1500 other) tape slots/yr. (5%/yr)	< 20% unused	• No action if there is sufficient capacity in blank tapes. • Migrate data to denser media (long term action) • Request squeezing and recycling of tapes from the recyclable candidate lists • Purge obsolete files from tapes and recycle them • Eject obsolete tape media whose data has been migrated to newer media • Begin the process to procure additional tape slots.
Blank Tapes	2 month blank supply.	3500 additional tapes/yr.	<=575 blanks	• Procure additional tapes. If budget permits, we try to purchase enough new tapes for approx. 1+ year. • Request squeezing and recycling of tapes from the recyclable candidate lists • Purge obsolete files from tapes and recycle them • For unanticipated increase in where getting delivery of new tapes in time is risky, temporarily borrow tapes from other customers.

Tape Drive Capacity	Current = 76 drives ~ 19 GB/s (+ some LTO4)	5%/yr	75% Utilization	We monitor tape-drive usage hours. If this exceeds about 75% capacity on average and continues to climb we will consider taking action. There are several strategies if the threshold is exceeded (all of which have been exercised at one time or the other): • Purchase additional tape drives. Each library can host 64 tape drives (a total of 448 for 7 libraries. Currently we have about 120 drives) • Move underutilized drives from one customer to the customer that is having a higher demand than usual. • Slow down competing migration processing freeing up drive resources • Work with customers to eliminate any inefficiencies in drive utilization.
Public Production Read/Write dCache capacity	100—200 day file lifetime	2560 TB, 100 days	< 90 day lifetime	1. Allocate more disk from reserve if available 2. See if some storage can be moved off to smaller dedicated read/write disk storage 3. Begin the process of purchasing more disk capacity.
Public Scratch dCache capacity	30 day file lifetime	1000 TB, 35 days	< 30 day lifetime	1. Allocate more disk from reserve if available 2. Begin the process of purchasing more disk capacity.
Public Analysis dCache capacity	Customer specific	1355 TB total	Customer specific	This is customer managed space. It is not a cache. If the customer needs more persistent space it makes request through the SCPPM. There are the following options: 1. Allocate more disk from reserve if available 2. Begin the process of purchasing more disk capacity

Table 1. Overall Capacity Management

Descriptions for each item in Table 1. Are described in the next section.

DETAILED PROCEDURES

Scientific Data Storage and Access service offerings

SDSA services are built from two service offerings, Enstore and dCache.

Enstore is an in-house developed tape management software application that manages capacity and access to the tapes and tape drives in automated tape libraries. dCache is a distributed disk cache storage system that acts as a front-end to Enstore, and can provide fast access to frequently accessed files. dCache is a collaboration between Fermilab, DESY, and NGDF. These two offerings share a common namespace called Chimera, which customers use to access their files.

There are several instances of Enstore and dCache at Fermilab:

1. CDF Enstore and dCache – CDF Tevatron Run II supported to 2020
2. D0 Enstore and dCache – D0 Tevatron Run II supported to 2020
3. Public Enstore and dCache – Intensity Frontier, Astro (DES, Darkside, etc.), LQCD
4. CMS Enstore and dCache (Enstore is part of the Public, but CMS has its own Chimera and dCache)
5. Active Archive Facility Enstore and dCache (A subset of the Public) for External customers. AAF has its own dedicated disk pools.

Tapes are stored in 3 Tape library complexes: A stand-alone "GS" 10,000 slot complex and a 30,000 slot "G1"complex in the Grid Computing Center, and a 30,000 slot "F1" complex in the Feynman Computing Center. Our customers are distributed across these complexes are as follows:

- F1: "Public" data: Intensity Frontier, DES, Darkside and other Astro, LQCD, Active Archive Facility, CDF and D0 reconstructed data
- GS: Second copies of subsets of Public Data, RMAN database backups, D0 and CDF RAW data
- G1: CMS data

Note that we do not "shelve" tapes, that is, we do not eject inactive tapes from the library and put them on a shelf. Except for loading and unloading tapes, and ejecting full tapes to physically write-protect them by flipping the write protect tab, tape storage is fully automated.

Disk cache storage and servers are all located in the Feynman Computing Center in the second and third floor computing rooms. There are 4 separate independent instances of dCache: Public, CMS, D0 and CDF:

- CMS: Tape backed and non-tape backed cache
- Public: Tape backed and non-tape backed cache and non-tape backed persistent storage.
- CDF: Tape backed cache
- D0: Tape backed cache

While Table 1. Applies across all of these instances, each instance is treated independently with respect to capacity management thresholds, with the exception that resources may be moved around between the instances to handle peaks or unanticipated needs.

Enstore Service Offering definitions

Tape slots are cells in the tape libraries that hold individual tapes. Once the number of empty slots (slots with no tape) falls below 20% we decide on whether any action is required. Note that there can be 0% empty slots but plenty of capacity if many of the occupied slots hold blank tapes. The possible actions are listed in the table. Empty tape slots represent potential capacity since blank tapes can be stored in the slots.

Blank Tapes are tape cartridges that do not yet have data written to them. Blank tapes represent real capacity. Since the cycle for procurement of tapes to loading them into the library can take up to 2 months, we typically want at least 2 months supply available for all customers.

Migrated tapes are the original tapes whose data has been migrated to newer media. Migrated tapes still hold a copy of the original data, but it is not active in the storage system (files are marked deleted), and are usually ejected

and sold or destroyed according to policy. Migrated tapes are potential capacity since they may be ejected to free up tape slots to which blank tapes can be stored.

Tape Drive's have fixed bandwidth capacity. Each drive can move data on a tape cartridge to and from dCache via Enstore mover computers. Public Enstore/dCache users share tape drives while CMS and Run II have their own. Equally (or more) important than bandwidth is tape drive utilization, measured in tape-drive hours / day. Poor usage patterns can lead to slow transfers from tape tying up tape drive resources that other customers may need. We monitor this activity and the possible actions we take when near capacity are listed in Table 1.

dCache Service Offering definitions

dCache production read/write storage is a tape backed cache shared by Public users. It acts as a cache in that least recently used (LRU) files are evicted from the cache to make room for new files. Evicted files are not lost since they are tape backed and a requested file that is not in this cache will automatically be fetched from tape to cache then transferred to the user. New files written to the cache remain resident until written to tape, after which they are eligible for LRU eviction. This disk cache has two related characteristics: capacity and file lifetime. File lifetime depends on the total capacity of the cache and the rate that files are written into the cache. We monitor the lifetime of the cache and the capacity used by each customer. The desired lifetime is be between 100 and 200 days – long enough for a production run to process datasets in the cache.

dCache scratch cache is a LRU cache that is not tape backed. If a file is evicted from this cache, it is not recoverable. Scratch storage is intended for short term intermediary file output from production jobs. The desired file lifetime for this cache is > 30 days.

dCache persistent storage is not tape backed and does not act as a cache – files are not automatically evicted. We provide customers with a fixed amount of dedicated persistent storage by request to the SCPPM. The customers manage the storage space – if it fills up, the customers must manually remove files to make room for more. We provide monitoring tools so each customer that has persistent storage can view how much space is used and who is using it.

Procedures

Decision making on capacity is described in detail Table 1. The trending and monitoring sections below describe the information that is used to trigger actions and to make decisions on what course of action should be taken. Most monitoring is performed on a daily basis by service administrators and are reviewed when generating data for monthly capacity reporting.

The capacity thresholds in Table 1. are conservative and based on past experience. Burn rate is monitored and projected into the future to determine if adjustments to the initial estimates may be needed. For dCache, file lifetimes are monitored and, if they are too short, more disk space may be acquired or disk allocation redistributed (for instance for peak demand by an experiment). The input data needed to determine if performance needs are met and for input to making decisions in Table 1 are described below.

CAPACITY AND PERFORMANCE REQUIREMENTS

As part of the SCPPM budget planning process each year, estimated capacity requirements are determined for each customer of the SDSA services, and the Scientific Data Storage and Access budget is constructed from these requirements. In addition to this planning with customers, new and emerging technology is evaluated annually.

Staffing resources also need to be considered for for this service. Staffing levels will be reviewed, reported, and updated yearly in the *Tactical Plan for Scientific Data Storage and Access* available at

https://tpa.fnal.gov/?SPHostUrl=https://fermipoint.fnal.gov/organization/cs/ocio/fm&SPLanguage=en-US&SPClientTag=21&SPProductNumber=15.0.4569.1000

TRENDING AND PREDICTIVE ANALYSIS

Tape usage trending is monitored by historical tape "burn" rate plots. These plots are organized per tape library and per media type within the library as well as per customer. For the "Public" storage for example:

http://www-stken.fnal.gov/enstore/all_sg_burn_rates.html

These plots show historical bytes written, blank tapes drawn in the last month and week, and remaining tapes. These are monitored by administrators on a daily basis. In addition, point in time tape capacity usage per media type and library are generated on a monthly basis for operations reporting and are also included in the quarterly reports. These are compared to the customers' capacity requests.

Tape drive utilization trending is monitored per library and media type and per customer. These are monitored on a daily basis by the administrators and on a monthly basis by the service owner.

dCache capacity usage and lifetimes are monitored by administrators on a monthly basis and are available in Monthly (operations) and Quarterly reports. An example with all customers stacked:

http://www-stken.fnal.gov/enstore/drive-hours/plot_enstore_system.html

Most trending plots are reported at the first Computing Sector operations meeting of each month, so are reviewed more frequently than quarterly. These monthly reports are available in the "Usage and Capacity charts" document in the document set:

https://cd-docdb.fnal.gov:440/cgi-bin/ShowDocument?docid=2478

MONITORING AND REPORTING

The data monitored for managing tape and tape drive capacity is:

- Free tape slots, active tape slots, blank tapes, recyclable tapes list per library complex
- Historical tape/capacity consumption trends (per customer and total)
- Historical tape drive utilization plots (per customer and total)
- Expected capacity needs for each customer from the SCPMT rollup.

and for dCache storage:

- Scratch file lifetime

- Production read/write file lifetime
- Per customer cache utilization
- Per customer bandwidth utilization
- Expected Capacity needs for each customer from the SCPMT rollup (especially for persistent space)

The services themselves are complex distributed systems and have a host of other monitoring to ensure they are operating properly.

Capacity and usage are reported at weekly Computing Sector operations meeting and are updated once/month (see link in previous section). In addition, quarterly reports for the services, which include capacity usage and trends are provided

https://cd-docdb.fnal.gov:440/cgi-bin/ShowDocument?docid=5517

GOVERNANCE

Capacity Manager and Service Level Manager will ensure that procedures are being followed and that the Service is meeting capacity commitments.

This procedure will be reviewed annually for accuracy and necessity

RELATED DOCUMENTS

Service Level Agreement and Commitments (docdb CS-doc-5032)

Capacity Plan for Core IT Services (docdb CS-doc-4047)

VERSION HISTORY

Version	Implemented By	Revision Date	Approved By	Approval Date	Reason
1.0	Gene Oleynik	10/15/2015	Brian Mckittrick	10/19/2015	

Next revision date: TBD